Density Estimation [Non-parametric]

$\therefore P(\text{points in } R) = \frac{k}{N} = \frac{4}{13}$

let V be volume of R, then for $\underline{x}$ in R:

$$P(\underline{x}) = \frac{\frac{k}{N}}{V}$$

$\frac{k}{N}$ → prob in a region

V → area of a region

For better estimation

1. $V \to 0$
2. $k \to \infty$
3. $\frac{k}{n} \to 0$

Method 1: Kernel Density Estimation (KDE)

- Create a window or kernel function $\Delta(\underline{x})$
  e.g. Parzen Window
- Center the function at each points $\Delta(\underline{x} - \underline{x}_i)$
- Take sum of all functions

$$P_n(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} \Delta(\underline{x} - \underline{x}_i)$$

We can define $\Delta(\underline{x} - \underline{x}_i)$ as $\frac{1}{V_n} k\left(\frac{\underline{x} - \underline{x}_i}{h}\right)$, choose V_n to ensure

$$\int \frac{1}{V_n} k\left(\frac{\underline{x} - \underline{x}_i}{h}\right) d\underline{x} = 1.$$

Then $P_n(\underline{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}k\left(\frac{\underline{x}-\underline{x}_i}{h}\right)$

Usually, V_n could be h, then:

h is width
($\underline{x}_i$ is within this width)

$$\boxed{P_n(\underline{x}) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{\underline{x}-\underline{x}_i}{h}\right)}$$

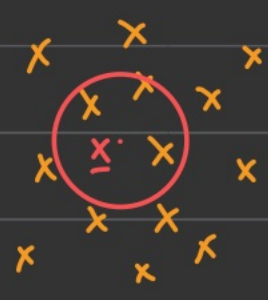
Method 2: K-Nearest Neighbors (KNN)

No width (h) parameter
No window/kernel

use $P(\underline{x}) = \frac{\frac{K}{N}}{V}$

- Choose k as a constant
- V is calculated based on k

let $k=3$, $N=15$



Now V may be 13

$P(\underline{x}) = \frac{\frac{3}{15}}{13}$

$= \frac{1}{65}$

Classification with Density Estimation [Generative Approach]

1. Estimate $P(\underline{x}|S_k)$ separately for each class S_k. [only use points in S_k]

2. Use or estimate $P(S_k)$. e.g. $\hat{P}(S_k) = \frac{n^{(k)}}{N}$ → points # in S_k / → total points #

3. Use Bayes Classifier. [min-error or min-risk]

$$\hat{P}(\underline{x}|S_k)\hat{P}(S_k) > \hat{P}(\underline{x}|S_j)\hat{P}(S_j) \quad \forall j \neq k$$

$$\Rightarrow \underline{x} \in S_k$$

Parameter Estimation [Classification Case]

Instead of estimate $P(\underline{x}|s_k)$ with KDE or KNN,

assume $P(\underline{x}|s_k) = f(\underline{x}, \underline{\theta}_k)$ ← known or assumed.
← calculated with points in s_k

Ex: $p(\underline{x}|s_k) = N(\underline{x}, \underline{m}_k, \underline{\underline{\Sigma}}_k)$
← unknown

① Consider $\underline{\theta}$ as deterministic

MLE (Maximum Likelihood Estimation)

Ex: $p(\underline{x}|\underline{\theta}) = N(\underline{x}, \underline{\mu}, \underline{\underline{\Sigma}})$ ← known
← unknown
← multivariate Gaussian

$$\therefore \hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} p(\underline{x}_1, \cdots, \underline{x}_n | \underline{\theta}) \Rightarrow L = \ln p(\underline{x}_1, \cdots, \underline{x}_n | \underline{\theta}) = \ln \prod p(\underline{x}_k|\theta) = \sum \ln p(\underline{x}_k|\theta)$$

$$\therefore \hat{\underline{\mu}}_{ML} = \arg\max_{\underline{\mu}} L$$

$$\nabla_{\underline{\mu}} L = \nabla_{\underline{\mu}} \sum \ln \frac{1}{\sqrt{(2\pi)^D |\underline{\underline{\Sigma}}|}} e^{-\frac{1}{2}(\underline{x}_k - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}_k - \underline{\mu})} = \nabla_{\underline{\mu}} \sum \ln \frac{1}{\sqrt{(2\pi)^D |\underline{\underline{\Sigma}}|}} - \frac{1}{2}(\underline{x}_k^T \underline{\underline{\Sigma}}^{-1} \underline{x}_k - \underline{x}_k^T \underline{\underline{\Sigma}}^{-1} \underline{\mu} - \underline{\mu}^T \underline{\underline{\Sigma}}^{-1} \underline{x}_k + \underline{\mu}^T \underline{\underline{\Sigma}}^{-1} \underline{\mu}) = 0$$

$$\Rightarrow \sum_{k=1}^{n} (-\underline{x}_k \underline{\underline{\Sigma}}^{-1} - \underline{\underline{\Sigma}}^{-1} \underline{x}_k + \underline{\underline{\Sigma}}^{-1} \hat{\underline{\mu}}_{ML} + \hat{\underline{\mu}}_{ML} \underline{\underline{\Sigma}}^{-1}) = 0$$

$$\sum_{k=1}^{n} (\underline{\underline{\Sigma}}^{-1} \underline{x}_k - \underline{\underline{\Sigma}}^{-1} \hat{\underline{\mu}}_{ML}) = 0$$

$$\boxed{\hat{\underline{\mu}}_{ML} = \frac{1}{n} \sum_{k=1}^{n} \underline{x}_k}$$

Now $P(\underline{x}|s_k) = N(\underline{x}, \hat{\underline{\mu}}_{ML}, \underline{\underline{\Sigma}})$

Similarly, if $\underline{\underline{\Sigma}}$ is not known

$$\nabla_{\underline{\underline{\Sigma}}} L = \nabla_{\underline{\underline{\Sigma}}} \sum_K \left[\ln \frac{1}{\sqrt{(2\pi)^D |\underline{\underline{\Sigma}}|}} - \frac{1}{2} (\underline{x}_K - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}_K - \underline{\mu}) \right] = 0$$

$$\ln \frac{1}{\sqrt{(2\pi)^D}} - \frac{1}{2} \ln |\underline{\underline{\Sigma}}|$$

$$= \sum_K \left[-\frac{1}{2} \underline{\underline{\Sigma}}^{-1} + \frac{1}{2} (\underline{x}_K - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} \underline{\underline{\Sigma}}^{-1} (\underline{x}_K - \underline{\mu}) \right] = 0$$

$$\Rightarrow \sum_K \left[-\hat{\underline{\underline{\Sigma}}} + (\underline{x}_K - \underline{\mu})(\underline{x}_K - \underline{\mu})^T \right] = 0$$

$$\sum_{k=1}^{n} (\underline{x}_K - \underline{\mu})(\underline{x}_K - \underline{\mu})^T = n \cdot \hat{\underline{\underline{\Sigma}}}$$

$$\Rightarrow \boxed{\hat{\underline{\underline{\Sigma}}}_{ML} = \frac{1}{n} \sum_{k=1}^{n} (\underline{x}_K - \underline{\mu})(\underline{x}_K - \underline{\mu})^T}$$ where $\underline{\mu} = \hat{\underline{\mu}}_{ML}$ or prior

Then $P(\underline{x} | S_k) = N(\underline{x}, \hat{\underline{\mu}}_{ML}, \hat{\underline{\underline{\Sigma}}}_{ML})$

Note: $\frac{\partial \underline{\underline{X}}^{-1}}{\partial \underline{\underline{X}}} = -\underline{\underline{X}}^{-1} \underline{\underline{X}}^{-1}$ ①

$\frac{\partial \ln |\underline{\underline{X}}|}{\partial \underline{\underline{X}}} = \underline{\underline{X}}^{-1}$ if $\underline{\underline{X}}$ symmetric ②

$\frac{\partial \underline{A}^T \underline{\underline{X}} \underline{A}}{\partial \underline{\underline{X}}} = \underline{A} \underline{A}^T$ if $\underline{\underline{X}}$ symmetric ③

② Consider $\underline{\theta}$ is random

MAP (Maximum A Posterior estimation)

$P(\underline{\theta}|\underline{z})$: posterior density of $\underline{\theta}$.

$$P(\underline{\theta}|\underline{z}) = \frac{P(\underline{z}|\underline{\theta}) \cdot P(\underline{\theta})}{P(\underline{z})}$$

known or assumed like in MLE

$$L = \ln P(\underline{\theta}|\underline{z}) = \ln P(\underline{z}|\underline{\theta}) + \ln P(\underline{\theta}) - \ln P(\underline{z}) \leftarrow \text{constant}$$

prior on $\underline{\theta}$

Similarly $\hat{\underline{\theta}}_{MAP} = \arg\max_{\underline{\theta}} L$

↳ Ex: $\underline{\theta}$ is size of tumors

$P(\underline{\theta})$ could be $N(\underline{\theta}, \underline{m}_\theta, \underline{\underline{\Sigma}}_\theta)$

according to experts, $\underline{m}_\theta = 8$ $\underline{\underline{\Sigma}}_\theta = 4$

with prior, $\hat{\underline{\theta}}_{MAP}$ could be solved.

Ex: suppose $P(\underline{x}|\underline{\theta}) = N(\underline{x}, \underline{\mu}, \underline{\underline{\Sigma}})$ (unknown)

$P(\underline{\mu}) = N(\underline{\mu}, \underline{m}_\mu, \underline{\underline{\Sigma}}_\mu)$ known, prior

$$L = \overbrace{\ln P(\underline{x}_1, \cdots, \underline{x}_n|\underline{\theta})}^{\text{same as MLE}} + \overbrace{\ln P(\underline{\theta})}^{\text{prior}} = \left[\sum_k \ln P(\underline{x}_k|\underline{\theta})\right] + \ln P(\underline{\theta}) \leftarrow P(\underline{\mu})$$

$$\therefore \nabla_{\underline{\mu}} L = \nabla_{\underline{\mu}} \left\{ \left[\sum_k \ln \frac{1}{\sqrt{(2\pi)^D |\underline{\underline{\Sigma}}|}} e^{-\frac{1}{2}(\underline{x}_k - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}_k - \underline{\mu})}\right] + \ln \frac{1}{\sqrt{(2\pi)^D |\underline{\underline{\Sigma}}_\mu|}} e^{-\frac{1}{2}(\underline{\mu} - \underline{m}_\mu)^T \underline{\underline{\Sigma}}_\mu^{-1} (\underline{\mu} - \underline{m}_\mu)} \right\}$$

$$= \sum_k -\frac{1}{2}\left(-(\underline{\underline{\Sigma}}^{-1})^T \underline{x}_k - \underline{\underline{\Sigma}}^{-1} \underline{x}_k + \underline{\underline{\Sigma}}^{-1} \underline{\mu} + (\underline{\underline{\Sigma}}^{-1})^T \underline{\mu}\right) - \frac{1}{2} \nabla_{\underline{\mu}} \left(\underline{\mu}^T \underline{\underline{\Sigma}}_\mu^{-1} \underline{\mu} - \underline{\mu}^T \underline{\underline{\Sigma}}_\mu^{-1} \underline{m}_\mu - \underline{m}_\mu^T \underline{\underline{\Sigma}}_\mu^{-1} \underline{\mu} + \underline{m}_\mu^T \underline{\underline{\Sigma}}_\mu^{-1} \underline{m}_\mu \right)$$

$\underline{\underline{\Sigma}}^{-1}$ is symmetric

$$= \sum_k -\frac{1}{2}\left(-2\underline{\underline{\Sigma}}^{-1} \underline{x}_k + 2\underline{\underline{\Sigma}}^{-1} \underline{\mu}\right) - \frac{1}{2}\left(2\underline{\underline{\Sigma}}_\mu^{-1} \underline{\mu} - 2\underline{\underline{\Sigma}}_\mu^{-1} \underline{m}_\mu\right) = 0$$

$$\Rightarrow \underline{\underline{\Sigma}}^{-1} \cdot \sum_{k=1}^n \underline{x}_k - n \cdot \underline{\underline{\Sigma}}^{-1} \cdot \hat{\underline{\mu}}_{MAP} = \underline{\underline{\Sigma}}_\mu^{-1} \cdot \hat{\underline{\mu}}_{MAP} - \underline{\underline{\Sigma}}_\mu^{-1} \underline{m}_\mu$$

$$\Rightarrow \boxed{(n \cdot \underline{\underline{\Sigma}}^{-1} + \underline{\underline{\Sigma}}_\mu^{-1}) \cdot \hat{\underline{\mu}}_{MAP} = \underline{\underline{\Sigma}}^{-1} \cdot \sum_{k=1}^n \underline{x}_k + \underline{\underline{\Sigma}}_\mu^{-1} \cdot \underline{m}_\mu}$$

where $\underline{\underline{\Sigma}}^{-1}$ is needed. ($\underline{\underline{\Sigma}}^{-1}$ could be $\underline{\underline{\Sigma}}_\mu^{-1}$)

Similarly

$$\nabla_{\Sigma} L = \nabla_{\Sigma}\left\{\left[\sum_k \ln \frac{1}{\text{Normalizer}\,|\Sigma|} e^{-\frac{1}{2}(x_k-\mu)^T \Sigma^{-1}(x_k-\mu)}\right] + P(\Sigma)\right\}$$

← could be matrix Gaussian

← out of my knowing area

Two approaches for $\underline{\theta}_k$ ← Ex. as in MLE

1) $P(\underline{x}|S_k) = N(\underline{x}, \underline{\mu}, \Sigma)$ $\quad \underline{\mu} = \hat{\underline{\mu}}_{MAP}$ (point estimate)

2) $P(\underline{x}|S_k) = \int N(\underline{x}, \underline{\mu}, \Sigma)\, P(\underline{\mu})\, d\underline{\mu}$

Bayes Classification

posterior density of $\underline{\mu}$

Then, use Bayes Classifier.

Parameter Estimation [Regression Case]

estimate $P(y|\underline{x}) = f(y, \underline{x}, \theta)$

Ex: $P(y|\underline{x}) = N(y, \hat{f}(\underline{x}, \underline{w}), \sigma_y^2)$ ← unknown; ← mean value

where $\hat{f}(\underline{x}, \underline{w})$ could be { linear: $\underline{w}^T \underline{x}$; nonlinear: $\underline{w}^T \phi(\underline{x})$

Maximum Likelihood Estimation (MLE)

$$p(\text{data}|\underline{\theta}) = P(\underline{\underline{x}}, \underline{y}|\underline{\theta}) = P(\underline{y}|\underline{\underline{x}}, \underline{\theta}) \cdot P(\underline{\underline{x}}|\theta)$$

↖ indep.

$$= P(\underline{y}|\underline{\underline{x}}, \underline{\theta}) \cdot P(\underline{\underline{x}})$$

$$\therefore L = \ln[P(\underline{y}|\underline{\underline{x}}, \underline{\theta}) P(\underline{\underline{x}})] = \ln P(\underline{y}|\underline{\underline{x}}, \underline{\theta}) + \ln P(\underline{\underline{x}})$$

$$\therefore \hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} L$$

$$L = \ln P(\underline{y}|\underline{\underline{x}}, \underline{\theta}) + \ln P(\underline{\underline{x}}) = \ln \prod_{i=1}^{n} P(y_i|\underline{x}_i, \underline{\theta}) + \ln \prod_{i=1}^{n} P(\underline{x}_i)$$

$$= \sum_{i=1}^{n} \ln P(y_i|\underline{x}_i, \underline{\theta}) + \sum_{i=1}^{n} \ln P(\underline{x}_i)$$

↙ assume given

$$\nabla_{\underline{\theta}} L = \nabla_{\underline{\theta}} \left(\sum_{i=1}^{n} \ln P(y_i|\underline{x}_i, \underline{\theta}) \right) = \nabla_{\underline{\theta}} \left(\sum_{i=1}^{n} \ln N(y_i, \hat{f}(\underline{x}_i, \underline{\theta}), \sigma_y^2) \right)$$

$$= \nabla_{\underline{\theta}} \left(\sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}\left(\frac{y_i - \hat{f}(\underline{x}_i, \underline{\theta})}{\sigma_y}\right)^2} \right) = \nabla_{\underline{\theta}} \left(\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i - \hat{f}(\underline{x}_i, \underline{\theta})}{\sigma_y}\right)^2 \right)$$

$$= + \sum_{i=1}^{n} \left(\frac{y_i - \hat{\underline{\theta}}^T \underline{x}_i}{\sigma_y} \right) \cdot \underline{x}_i / \sigma_y = 0$$

$$\therefore \sum_{i=1}^{n} (y_i - \hat{\underline{\theta}}^T \underline{x}_i) \underline{x}_i = 0 = (\underline{y} - \underline{\underline{x}} \hat{\underline{\theta}})^T \cdot \underline{\underline{x}} = \underline{y}^T \underline{\underline{x}} - \hat{\underline{\theta}}^T \underline{\underline{x}}^T \cdot \underline{\underline{x}} = \underline{0}$$

$$\Rightarrow \quad \underline{\underline{X}}^T \underline{Y} = \underline{\underline{X}}^T \underline{\underline{X}}\, \hat{\underline{\theta}} \quad \Rightarrow \quad \hat{\underline{\theta}}_{ML} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{Y}$$

$$\therefore \ \hat{\underline{w}}_{ML} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{Y} = \underline{\underline{X}}^{-} \underline{Y} \quad \text{[linear regression, Gaussian pdf, MLE]}$$

← same as OLS

$$\nabla_{\sigma_Y} L = \nabla_{\sigma_Y} \left(\sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_Y - \frac{1}{2} \cdot \left(\frac{Y_i - \underline{\theta}^T \underline{x}_i}{\sigma_Y} \right)^2 \right) = 0$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{\sigma_Y} + \left(\frac{Y_i - \underline{\theta}^T \underline{x}_i}{\sigma_Y} \right) \cdot \frac{Y_i - \underline{\theta}^T \underline{x}_i}{\sigma_Y^2} \right]$$

$$= -\frac{n}{\sigma_Y} + \sum_{i=1}^{n} \frac{(Y_i - \underline{\theta}^T \underline{x}_i)^2}{\sigma_Y^3}$$

$$\Rightarrow \quad \sum_{i=1}^{n} (Y_i - \underline{\theta}^T \underline{x}_i)^2 = n \cdot \sigma_Y^2$$

$$\Rightarrow \quad \boxed{\hat{\sigma}_{Y_{ML}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \underline{\theta}^T \underline{x}_i)^2}{n}}} \quad \text{where } \underline{\theta} = \hat{\underline{w}}_{ML}$$

Then $P(y|\underline{x}) = N(y, \ \hat{\underline{w}}_{ML} \underline{x}, \ \hat{\sigma}^2_{Y_{ML}})$

Maximum A Posterior (MAP)

$$\hat{\underline{\theta}}_{MAP} = \arg\max_{\underline{\theta}} L = \arg\max_{\underline{\theta}} \ln p(\underline{\theta} | data)$$

Similarly: $P(\underline{\theta} | data) = P(\underline{\theta} | \underline{\underline{X}}, \underline{Y}) = \dfrac{P(\underline{\underline{X}}, \underline{Y} | \underline{\theta}) P(\underline{\theta})}{P(\underline{\underline{X}}, \underline{Y})}$ ← constant for $\underline{\theta}$

$$\therefore L = \ln P(\underline{\underline{X}}, \underline{Y} | \underline{\theta}) P(\underline{\theta}) = \ln P(\underline{Y} | \underline{\underline{X}}, \underline{\theta}) \cdot P(\underline{\underline{X}} | \underline{\theta}) \cdot P(\underline{\theta})$$

← indep. & constant for $\underline{\theta}$

$$= \ln P(\underline{Y} | \underline{\underline{X}}, \underline{\theta}) \cdot P(\underline{\theta})$$

Similarly $P(\underline{Y} | \underline{\underline{X}}, \underline{\theta}) = \prod_{i=1}^{n} P(Y_i | \underline{\underline{X}}, \underline{\theta}) = \prod_{i=1}^{n} P(Y_i | \underline{X}_i, \underline{\theta})$

and $P(Y_i | \underline{X}_i, \underline{\theta}) = N(Y_i, f(\underline{X}_i, \underline{\theta}), \sigma_Y^2)$ ← assume given

and $P(\underline{\theta}) = N(\underline{\theta}, \underline{m}_\theta, \underline{\underline{\Sigma}}_\theta)$ ← known (prior)

where θ_i & θ_j, $i \neq j$, are indep. ; so $\underline{\underline{\Sigma}}_\theta = \sigma_\theta^2 \cdot \underline{\underline{I}}$

$$\Rightarrow \nabla_{\underline{\theta}} L = \nabla_{\underline{\theta}} \left(\ln \prod_{i=1}^{n} \underbrace{N(Y_i, f(\underline{X}_i, \underline{\theta}), \sigma_Y^2)}_{\frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{Y_i - f(\underline{X}_i, \underline{\theta})}{\sigma_Y}\right)^2}} + \ln \underline{P(\underline{\theta})} \right)$$

$$= +\sum_{i=1}^{n} \left(\frac{Y_i - \hat{\underline{\theta}}^T \underline{X}_i}{\sigma_Y} \right) \cdot \underline{X}_i / \sigma_Y + \nabla_{\underline{\theta}} \left(\ln \frac{1}{(2\pi)^{\frac{D}{2}} |\underline{\underline{\Sigma}}_\theta|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\underline{\hat{\theta}} - \underline{m}_\theta)^T \underline{\underline{\Sigma}}_\theta^{-1} (\underline{\hat{\theta}} - \underline{m}_\theta) \right]} \right)$$

$$= \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (Y_i - \hat{\underline{\theta}}^T \underline{X}_i) \cdot \underline{X}_i + -\frac{1}{2\sigma_\theta^2} \left[\underline{\hat{\theta}} + \underline{\hat{\theta}} - \underline{m}_\theta - \underline{m}_\theta \right]$$

$\underline{\theta}^T \underline{\theta} - \underline{\theta}^T \underline{m}_\theta - \underline{m}_\theta^T \underline{\theta} - \underline{m}_\theta^T \underline{m}_\theta$

$$= 0$$

$$\Rightarrow \quad \frac{1}{\sigma_\theta^2}(\underline{\theta}-\underline{m}_\theta) = \frac{1}{\sigma_y^2}\sum_{i=1}^{n}(y_i - \underline{\theta}^T\underline{x}_i)\cdot\underline{x}_i = \frac{1}{\sigma_y^2}\left[(\underline{y}-\underline{\underline{x}}\,\underline{\theta})^T\underline{\underline{x}}\right]^T = \frac{1}{\sigma_y^2}\underline{\underline{x}}^T(\underline{y}-\underline{\underline{x}}\,\underline{\theta})$$

$$\Rightarrow \quad \frac{1}{\sigma_\theta^2}\underline{\theta} + \frac{1}{\sigma_y^2}\underline{\underline{x}}^T\underline{\underline{x}}\,\underline{\theta} = \frac{1}{\sigma_0^2}\underline{m}_\theta + \frac{1}{\sigma_y^2}\underline{\underline{x}}^T\underline{y}$$

$$\left(\frac{\sigma_y^2}{\sigma_0^2}\cdot\underline{\underline{I}} + \underline{\underline{x}}^T\underline{\underline{x}}\right)\underline{\theta} = \frac{\sigma_y^2}{\sigma_\theta^2}\underline{m}_\theta + \underline{\underline{x}}^T\underline{y}$$

$$\Rightarrow \quad \underline{w}_{MAP} = \underline{\theta}_{MAP} = \left(\underline{\underline{x}}^T\underline{\underline{x}} + \frac{\sigma_y^2}{\sigma_w^2}\underline{\underline{I}}\right)^{-1}\left(\underline{\underline{x}}^T\underline{y} + \frac{\sigma_y^2}{\sigma_w^2}\underline{m}_w\right)$$

Similarly, Two approaches for $\underline{w}$

1) $P(y|\underline{x}) = N(y,\ f(\underline{x}, \underline{\hat{w}}_{MAP}),\ \sigma_y^2)$ (point estimate)

2) $P(y|\underline{x}) = \int N(y, f(\underline{x}, \underline{w}), \sigma_y^2)\, P(\underline{w})\, d\underline{w}$

Bayes Linear Regression ↗

↖ need to find by $P(\underline{w}|\text{data})$

Predict $y(\underline{x})$

$$\hat{y}(\underline{x}) = E[y|\underline{x}] = \int y\, P(y|\underline{x})\, dy$$

We have $P(y|\underline{x}) = N(y, \underline{f(\underline{x}, \underline{\hat{w}})}, \sigma_y^2)$ ← expectation

$$\therefore \hat{y}(\underline{x}) = f(\underline{x}, \underline{\hat{w}}) = \underline{\hat{w}}^T\underline{x} \begin{cases} \underline{\hat{w}}_{ML}\,\underline{x} \\ \underline{\hat{w}}_{MAP}\,\underline{x} \end{cases}$$

Bayesian Linear Regression

In MAP, we maximize $P(\underline{\theta}|\text{data})$ and find $\hat{\theta}_{MAP}$

Now, we integrate over it ← $P(\underline{\theta}|\text{data})$

$$P(y|\underline{x}) = \int N(y, f(\underline{x}, \underline{w}), \sigma_y^2)\, P(\underline{w})\, d\underline{w}$$

$$P(\underline{\theta}|\text{data}) = P(\underline{\theta}|\underline{\underline{X}}, \underline{y}) = \frac{P(\underline{\underline{X}}, \underline{y}|\underline{\theta})P(\underline{\theta})}{P(\underline{\underline{X}}, \underline{y})} = \frac{P(\underline{y}|\underline{\underline{X}}, \underline{\theta}) \cdot P(\underline{\underline{X}}|\underline{\theta}) \cdot P(\underline{\theta})}{P(\underline{\underline{X}}, \underline{y})}$$

$P(\underline{\underline{X}})$ because data is independent of $\underline{\theta}$

$$= K \cdot P(\underline{y}|\underline{\underline{X}}, \underline{\theta}) \cdot P(\underline{\theta})$$

K: constant; $P(\underline{\underline{X}}, \underline{y})$: constant for $\underline{\theta}$

Similarly, $P(y_i|\underline{x}_i, \underline{\theta}) = N(y_i, f(\underline{x}_i, \underline{\theta}), \sigma_y^2)$ ← assume given

$$P(\underline{\theta}) = N(\underline{\theta}, \underline{m}_\theta, \underline{\underline{\Sigma}}_\theta)$$

known (prior)

where θ_i & θ_j, $i \neq j$, are indep.; so $\underline{\underline{\Sigma}}_\theta = \sigma_\theta^2 \cdot \underline{\underline{I}}$

$$\therefore P(\underline{y}|\underline{\underline{X}}, \underline{\theta}) = \prod_{i=1}^{n} P(y_i|\underline{x}_i, \underline{\theta}) = \prod_{i=1}^{n} P(y_i|\underline{x}_i, \underline{\theta}) = \prod_{i=1}^{n} N(y_i, f(\underline{x}_i, \theta), \sigma_y^2)$$

$$= N(\underline{y}, \underline{f}(\underline{\underline{X}}, \underline{\theta}), \underline{\underline{\Sigma}}_y) \quad \text{with } \underline{\underline{\Sigma}}_y = \sigma_y^2 \underline{\underline{I}}$$

Then $P(\underline{\theta}|\text{data}) = K \cdot N(\underline{y}, \underline{f}(\underline{\underline{X}}, \underline{\theta}), \underline{\underline{\Sigma}}_y) \cdot N(\underline{\theta}, \underline{m}_\theta, \underline{\underline{\Sigma}}_\theta)$

Omitting the algebra:

$$P(\underline{\theta}|\text{data}) = N(\underline{\theta}, \underline{m}_p, \underline{\underline{\Sigma}}_p)$$

with $\underline{m}_p = \underline{\underline{\Sigma}}_p \left(\underline{\underline{\Sigma}}_\theta^{-1} \underline{m}_\theta + \frac{1}{\sigma_y^2} \underline{\underline{X}}^T \underline{y} \right) = \underline{\underline{\Sigma}}_p \left(\frac{1}{\sigma_\theta^2} \underline{m}_\theta + \frac{1}{\sigma_y^2} \underline{\underline{X}}^T \underline{y} \right)$

$\underline{\underline{\Sigma}}_p^{-1} = \underline{\underline{\Sigma}}_\theta^{-1} + \frac{1}{\sigma_y^2} \underline{\underline{X}}^T \underline{\underline{X}} = \frac{1}{\sigma_\theta^2} \underline{\underline{I}} + \frac{1}{\sigma_y^2} \underline{\underline{X}}^T \underline{\underline{X}}$

↑ Sample correlation of feature with feature.

↑ Sample correlation of feature with output.

$\therefore P(y|\underline{x}) = \int N(y, f(\underline{x}, \underline{\theta}), \sigma_y^2) p(\underline{\theta}|data) d\underline{\theta}$

$= \int N(y, f(\underline{x}, \underline{\theta}), \sigma_y^2) N(\underline{\theta}, \underline{m}_p, \underline{\underline{\Sigma}}_p) d\underline{\theta}$

After significant algebra and integration

$\therefore P(y|\underline{x}) = N(y, \underline{m}_p^T \underline{x}, \sigma^2(\underline{x}))$

with $\sigma^2(\underline{x}) = \sigma_y^2 + \underline{x}^T \underline{\underline{\Sigma}}_p \underline{x}$ and $\underline{m}_p$, $\underline{\underline{\Sigma}}_p$ as before.

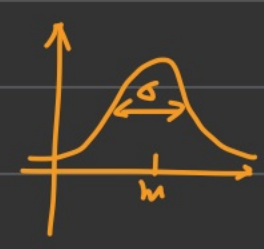
$\therefore \hat{\underline{w}} = \underline{m}_p \qquad \hat{y}(\underline{x}) = \underline{m}_p^T \cdot \underline{x}$

Nonlinear Regression [parameter estimation]

for all formula above: $\underline{x} \to \underline{\phi}(\underline{x})$, $\underline{\underline{X}} \to \underline{\underline{\Phi}}$

Ex. We can just set $\underline{\phi}(x) = e^{-\frac{(x-m)^2}{2\sigma^2}}$

- m: center
- σ: width



Summary of parameter estimation

Method	Classification	Regression
Theoretically optimal (know all pdf's)	Bayes min-error (or min-risk)	Min [TMSE] = min [E{square error}]
$\underline{\theta} = \underline{\hat{\theta}}_{ML}$	✓	✓
$\underline{\theta} = \underline{\hat{\theta}}_{MAP}$ (point estimate)	✓	✓
Bayesian classification / regression (integrate over $\underline{\theta}$ using $p(\underline{\theta}\|\text{data})$)	(not mentioned)	✓